わかりやすい日本語を設計できるか？

佐藤理史

名古屋大学工学研究科
電子情報システム専攻

© Satoshi Sato, 2006

概要

- 円滑な情報伝達を支援する日本語とはどのような日本語か？
  - そのような日本語を設計できるか？
- 円滑な情報伝達を支援する言語処理技術とはどのようなものか？
  - それは、どのようにしたら実現できるか？

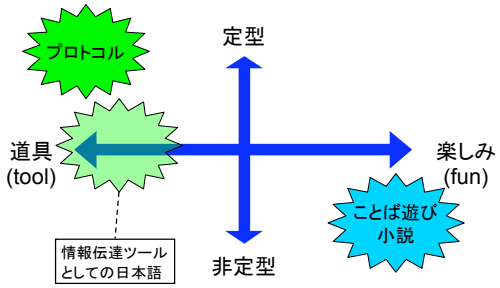
© Satoshi Sato, 2006

目次

- 総論
  - イントロダクション
  - 科研プロジェクトの概要
- 各論
  - 基本語彙を選定する

© Satoshi Sato, 2006

ことばとその機能



© Satoshi Sato, 2006

情報伝達ツールとしての日本語

- 緊急時の情報伝達（災害・テロ）
- インフォームド・コンセント （病気）
- 取扱い説明書やマニュアル（←PL法対策）
- 教育
  - 対象
    - 日本人（子供から老人まで）
    - 国内在住の外国人

情報がきちんと伝わることを
どうやって保証するか？
（テキストのquality control）
→
ある種の
controlled language
が必要

© Satoshi Sato, 2006

Questions

- 円滑な情報伝達を支援する日本語とはどのような日本語か？
  - そのような日本語を設計できるか？
- 円滑な情報伝達を支援する言語処理技術とはどのようなものか？
  - それは、どのようにしたら実現できるか？

© Satoshi Sato, 2006

方針

- 円滑な情報伝達を支援する日本語とはどのような日本語か？
  - そのような日本語を設計できるか？
- 円滑な情報伝達を支援する言語処理技術とはどのようなものか？
  - それは、どのようにしたら実現できるか？

→ わかりやすい日本語
日本語のサブセット

→ 規格として明確に定義

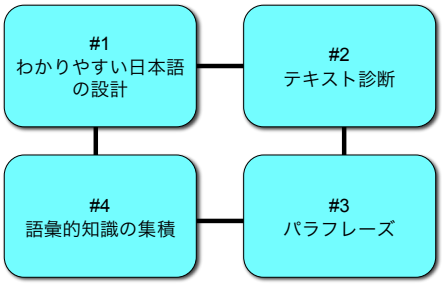
→ 「わかりやすい日本語」での文章作成支援
難易度判定とパラフレーズ支援

→ プロトタイプシステムの試作

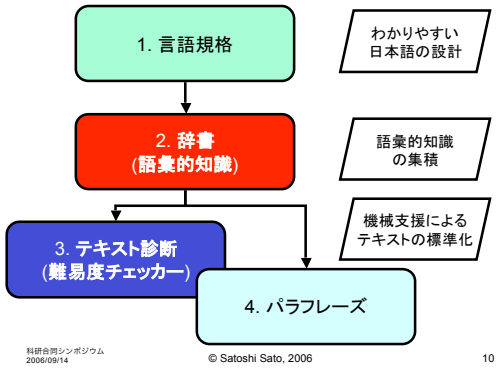
円滑な情報伝達を支援する言語規格と言語変換技術

- 科学研究費補助金(基盤研究A)
  - 2004年度から2007年度(4年間)
  - 研究費　総額3740万
- 研究代表者　佐藤理史(名大)
- 研究分担者
  - 宇津呂武人(筑波大)
  - 山本和英(長岡技科大)
  - 乾健太郎(奈良先端大)
  - 藤田篤(名大)

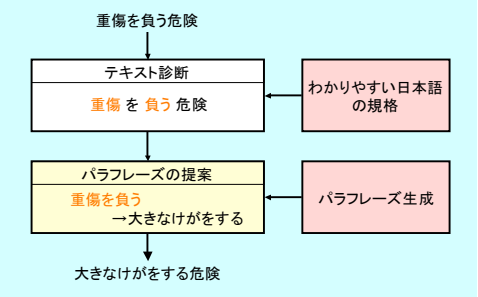
４つの柱



プロジェクトの全体像

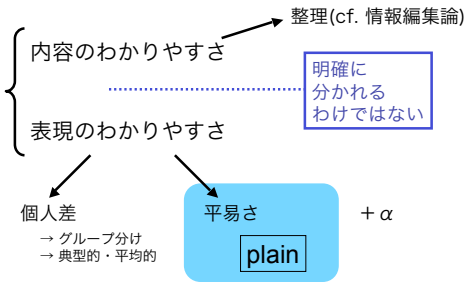


シナリオ



#1 わかりやすい日本語の設計

テキストのわかりやすさ



© Satoshi Sato, 2006

表現のコントロール

- マスコミ
  - 新聞
    - 記者ハンドブック - 新聞用字用語集 第10版 - (共同通信社)
    - 05-06版 朝日新聞社の用語の手引き
  - NHK
    - NHK 新用字用語辞典 第3版
- 役所
  - 外来語・役所ことば言い換え帳 (杉並区役所)
- 「やさしい日本語」（弘前大学　佐藤和之）
  - 外国人向け；災害情報の伝達
  - 日本語能力試験3級に準拠

© Satoshi Sato, 2006

日本語能力試験 出題基準 (LPT)

級分け
４級、３級、２級、１級

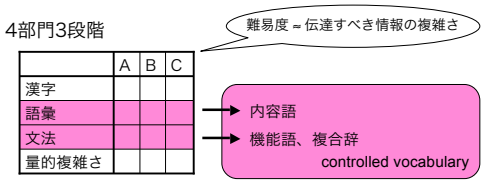
部門
文字、語彙、文法、
聴解、読解

初版1994、改訂版2002

© Satoshi Sato, 2006

平易な日本語を定義する

4部門3段階

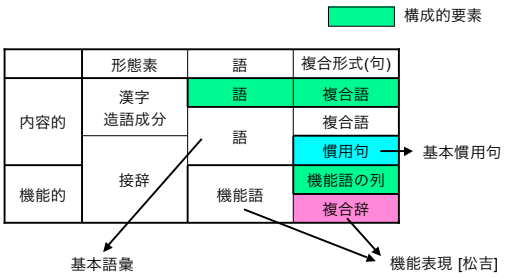


ユニットの階層構造

形態素 ＜ 語 ＜ 句 ＜ 節 ＜ 文

© Satoshi Sato, 2006

基本要素と構成的要素

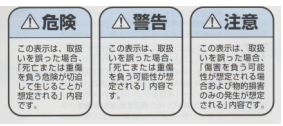


© Satoshi Sato, 2006

#2 テキスト診断

© Satoshi Sato, 2006

むずかしい表現？



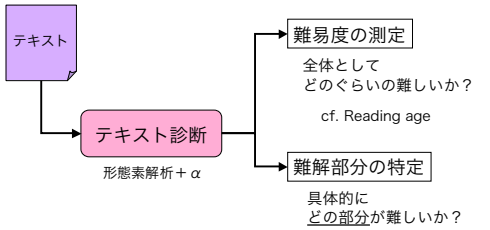
NTT DoCoMo ムーバN504i
取扱説明書 基本編

この表示は、取扱いを誤った場合、「死亡または重傷を負う危険が切迫して生じることが想定される」内容です。

人が死亡または重傷を負うおそれが高い内容を示しています。

使い方をまちがえると、死んだり大きなけがをしたりすることがらです。

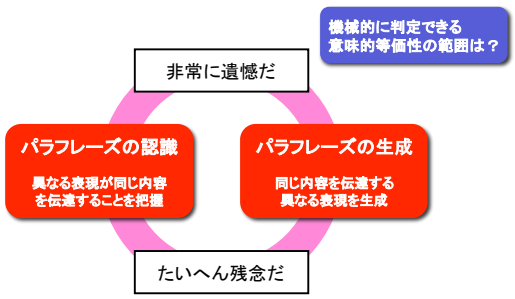
難易度の測定と難解部分の特定



cf. リーディング チュー太
(日本語読解学習支援システム)

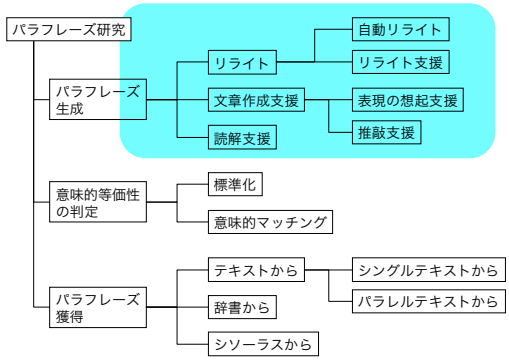
#3 パラフレーズ

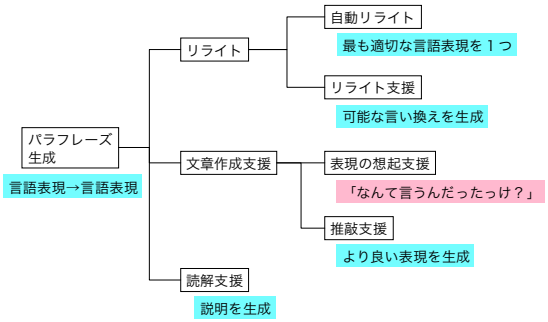
パラフレーズの研究



パラフレーズの応用領域

- テキストの利用者・状況適応
  - 情報弱者支援（情報バリアフリー）
    - 例：至急避難してください。
      →すぐに逃げなさい。
  - 携帯電話／PDA用テキストの自動生成
    - 例：文部省は昨日、大学への運営交付金を一律8%削減することを発表した。[34字]
      →大学への交付金8%削減を発表ー文部省 [18字]
- テキストの標準化と品質保証
- 機械翻訳／情報検索の高度化





パラフレーズ生成の実現

- 「語」の言い換え辞書
  - 静的に作る ← 差分情報
  - (ある種の)シソーラス

作成すべき基本データ

- 「句・節」の言い換え辞書
  - 動的=言い換え生成機構
  - Dynamic phrasal thesaurus

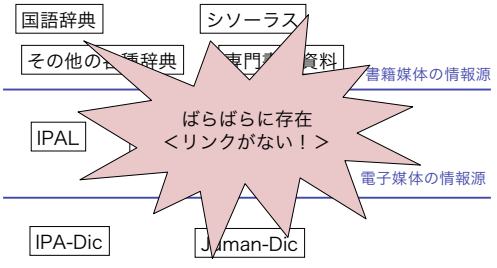
実現すべき基本機能

#4 語彙的知識の蓄積

必要な語彙的情報

- 代表表記
- よみ
- 可能な表記
- 語構成
- 品詞
- 活用等
  - 可能、受身、使役等の可否
- 意志性
- 意味
- 格フレーム
- コロケーション
- 派生語
  - 自他の対応
  - 他の品詞
- 類義語・反義語
- 言い換え

利用可能な情報源



辞書アライメント

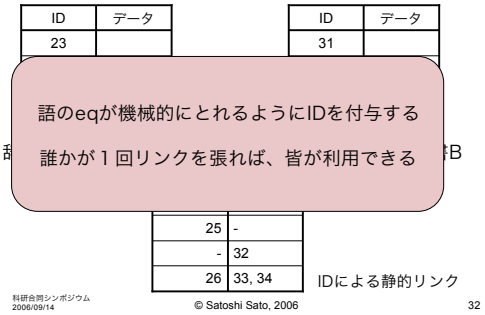
- 複数の辞書の見出し語集合間の対応を取る
  - 単純ではない
    - 語のeqが機械的に判定できない！
    - 人手のチェックが欠かせない
- 主な問題
  - 見出し語の不一致
  - 一語の範囲の不一致
  - 表記のゆれ
  - 品詞の不統一
  - 単位の不統一

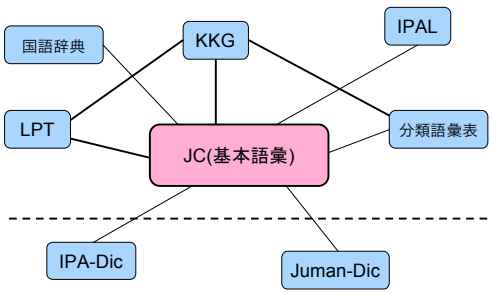
eg. きがえる⇔きかえる(着替える)
eg. さわ(触・障)る⇔触る, 障る
ろく⇔ろくな, ろくに
eg. カルタ⇔かるた
ロープウエー⇔ロープウェイ
形容動詞まわり

選択肢

- プログラムにより動的なリンクを張る
  - 見出し語の不統一、表記のゆれ、品詞の不統一等に対処しきれない
- すべてを包括する統合辞書を作成
  - 理想的だが、現実的ではない
- IDを用いて静的なリンクを張る
  - 現実的な解；段階的集積が可能

IDリンクによる語彙的資源の蓄積





基本語彙を選定する

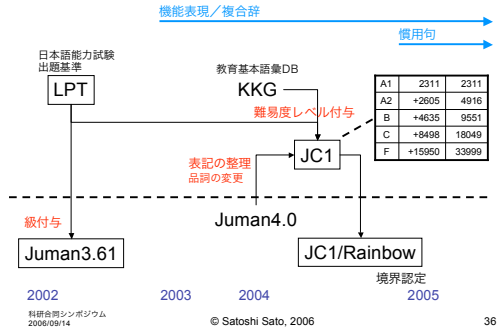
基本語彙を選定する

- 基本語彙
  - 使ってよい語彙 (controlled vocabulary)
- 基本設計

2500	
+2500	5000
+5000	10000
+10000	20000

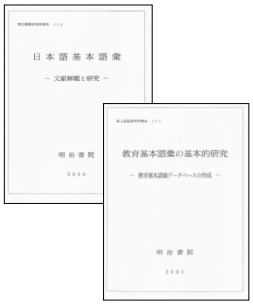
1. 基本語彙
   2000語内外が多い
2. 上位1万語で
   90%以上をカバー
3. 小学生用国語辞典
   2万から3万強

これまでの経過



基本語彙の基本文献

- 日本語基本語彙
  -- 文献解題と研究 --
  国立国語研究所報告116
  明治書院, 2000.
- 教育基本語彙の基本的研究
  --教育基本語彙データベースの作成-- [KKG]
  国立国語研究所報告117
  明治書院, 2001.
  - 7種類の基本語彙
  - 岩波国語辞典で正規化
  - CD-ROM付き



© Satoshi Sato, 2006

使用した資料 (1)

略称	出典	編著者名	発行年	収録語数	特記事項
LPT	日本語能力試験出題基準[改訂版]	国際交流基金・日本国際教育協会	2002年	7946語	日本語能力試験
国研 [KKG]	日本語教育のための基本語彙調査	国立国語研究所	1984年	6104語	外国人のための日本語教育
中央 [KKG]	中央教育基本語彙	中央教育研究所	1984年	4336語	小・中学校用

正規化した語数

© Satoshi Sato, 2006

KKG + LPT = KKG+

- LPTの校正
- LPTの正規化＝KKGへの統合
  - 27,234 + 325 = 27,559

級	公称	正規化
4	728	729
3	1409	1385
2	5035	4969
1	8009	7946

但し、あいさつ表現を除く

© Satoshi Sato, 2006

2500+2500を選ぶ

中央		国研		LPT		最少	和	差	JC
				4	729	729	729	0	
				3	1385	1385	1385	0	
A	1783					1783	2241	458	
		◎	2071			2071	2827	756	2695
B	3381					3381	3883	502	
C/X	4336					4336	4654	318	
				2	4969	4969	6025	1055	5554
		○	6104			6104	7280	1176	
				1	7946	7946	8819	872	

© Satoshi Sato, 2006

残り5000を選ぶ

LPT+国研+中央から
残り3265語

児言+池原+田中から
新たに1869語

→ 選択 → 4836語

略称	出典	編著者名	発行年	収録語数	特記事項
児言 [KKG]	言語要素指導	児童言語研究会	1962年	1843語	小・中学校用
池原 [KKG]	国語教育のための基本語体系	池原楢雄	1957年	2989語	小学校１～３年用
田中 [KKG]	学習基本語彙	田中久直	1956年	3456語	小学校用

© Satoshi Sato, 2006

JC2語彙表(2006/09/05版)

レベル	語数	累計
A	2695	
B	2859	5554
C	4836	10390

追加
- まじ
- 真横(まよこ)

LPTから除外
- 植わる
- 大部(たいぶ)

国研から除外
- 術語
- 末子(すえこ)
- 清音
- 促音
- 濁音
- 長音
- 録音テープ

© Satoshi Sato, 2006

今後の予定と主な問題

品詞別チェック	副詞の見出し語形 多品詞語　(eg. 現金 vs 現金な)
単位のチェック	造語成分・接辞の切り分け 複合名詞（どこまでを見出し語？） 複合動詞（どこまでを見出し語？） 連語
派生語チェック	どこまでを見出し語とするか？ eg. 形容詞+さ、動詞連用形
品詞体系の決定	形容動詞をどうするか？
形態素解析用辞書との アライメント	品詞体系の違い 可能な表記の選定

© Satoshi Sato, 2006

慣用句？

- 手に入れる　（入手する）
- 役に立つ　（有用だ）
- 頭がいい　（賢い）
- 気に入る
- 気になる
- 口が軽い　（おしゃべりだ）
- 口を出す　（意見を言う）

基本語彙にいれなくていいの？

© Satoshi Sato, 2006

慣用句の研究

- 2005年度
  - 慣用句用例データベースの試作 [尾島06]
  - 動詞慣用句の自動検出 [橋本06]
- 2006年度
  - 基本慣用句の選定
  - 基本慣用句用例データベースの作成

© Satoshi Sato, 2006

基本慣用句？

- 資料
  - 常用慣用句一覧[宮地 1982]
    - 学習用国語辞典５種類より収集＋α
    - 1280句
  - 新レインボー小学国語辞典 改訂第3版[2005]
    - 公称3万5千語
    - 慣用句、ことわざ、四字熟語、故事成句を明示
    - 1916句（こ344, 四262, 故63）

© Satoshi Sato, 2006

わかりやすい日本語を設計できるか？

© Satoshi Sato, 2006

現時点での答え

- controlled vocabulary でいく
  - 内容語　2500, 5000, 10000語
  - 機能表現
  - 慣用句　??

  } → JC2

- 上記のJC2がスタートポイント

戦っています！

© Satoshi Sato, 2006

Plain Japaneseの必要性

- 日本在住者の多様化
- 計算機とのコミュニケーション
  - ことばを話すロボット（日本語生成）
- 母国語能力の重要性
  - 国力の基礎を支える母国語能力
  - 外国語能力を支える母国語能力
    - 「外国語を身につけるための日本語レッスン」（三森ゆりか）

© Satoshi Sato, 2006